A MULTIMODAL APPROACH TO MUSIC TRANSCRIPTION

Marco Paleari, Benoit Huet

Multimedia Department
Eurecom Institute
Sophia Antipolis, France

Antony Schutz, Dirk Slock

Mobile Communication Department
Eurecom Institute
Sophia Antipolis, France

ABSTRACT

Music transcription refers to extraction of a human readable and interpretable description from a recording of a music performance. Automatic music transcription remains, nowadays, a challenging research problem when dealing with polyphonic sounds or when removing certain constraints.

Some instruments like guitars and violins add ambiguity to the problem as the same note can be played at different positions. When dealing with guitar music tablature are, often, preferred to the usual music score, as they present information in a more accessible way.

Here, we address this issue with a system which uses the visual modality to support traditional audio transcription techniques. The system is composed of four modules which have been implemented and evaluated: a system which tracks the position of the fretboard on a video stream, a system which automatically detects the position of the guitar on the first fret to initialize the first system, a system which detects the position of the hand on the guitar, and finally a system which fuses the visual and audio information to extract a tablature.

Results show that this kind of multimodal approach can easily disambiguate 89% of notes in a deterministic way.

Index Terms— Multimodal, Music Transcription, Guitar, Tablature

1. INTRODUCTION

Written music is traditionally presented as a score, a musical notation which includes attack times, duration and pitches of the notes that constitute the song.

When dealing with the guitar this task is usually more complex. In fact, the only pitch of the note is not always enough to represent the movements and the positions that the performer has to execute to play a piece. A guitar can indeed chime the same note (i.e. a note with the same pitch) at different positions of the fretboard on different strings (See Fig. 1). This is why the musical transcription of a guitar usually takes form of a tablature.

A tablature is a musical notation which includes six lines (one for each guitar string) and numbers representing the position at which the string has to be pressed to perform a note

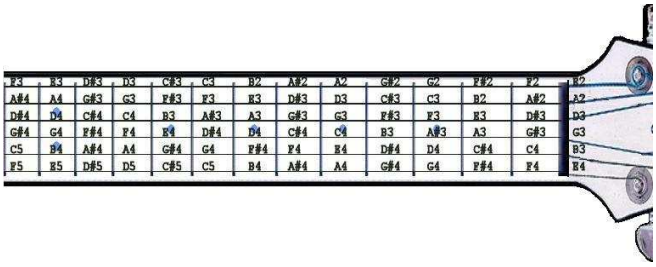


Fig. 1. Notes on a guitar fretboard

with a given pitch (figure 2). Special notations are added to represent particular effects like bend (‘^’), hammer on (‘h’), pull off (‘p’), etc.

The information about the movements which one player should execute to perform a piece can also be referred under the name of fingering. Burns and Wanderley [1] report few attempts that have been done to automatically extrapolate fingering information through computer algorithms:

1. real time processing using midi guitar
2. post processing using sound analysis
3. post processing using score analysis

Verner [2] retrieves fingering information through the use of midi guitar. Using a midi guitar with different midi channels associated to each different string, Verner can, in real time, extract the complete tablature. The study points out that users of midi guitars reported false note detections, difficulties while playing, and problems of synchronization among the different strings. In any case this approach is not always applicable because it needs expensive equipment which, on top of that, is not usually used by performers on scene.

Traube [3] suggests a solution based on the timbre of a guitar which only relies on the audio recording of a guitarist. Indeed, even if two notes have the same pitch they can have different timbre; with a-priori knowledge on the timbre of a guitar, it is therefore possible to estimate the fingering position. Common issues are precision, needs for a-priori knowledge, and monophonic operation limitation.

Another possibility is to analyze the produced score and to extract the tablature by applying a set of rules based on physical constraints of the instrument, biomechanical limitations, and others philological analysis. This kind of methods can result [4] in tablatures which are similar to the one generated by humans, but hardly deal with situations in which the artistic intention or skill limitations are more important than the biomechanical movement.

Last but not least, Burns and Wanderley [1] propose to use the visual modality to extract the fingering information. Their approach makes use of a camera mounted on the head of the guitar and extracts fingering information on the first 5 frets by using a circular Hough transformation to detect finger tips. Their system was positively evaluated in some preliminary studies but is not applicable to all cases because it needs ad hoc equipment, configuration, and it only returns information about the first 5 frets. Similarly Zhang et al. [5] track finger tips on a violin with a B-spline model of fingers contours.

This paper presents a multimodal approach to address this issue. The proposed approach combines information from video (webcam quality) and audio analysis in order to resolve ambiguous situations.

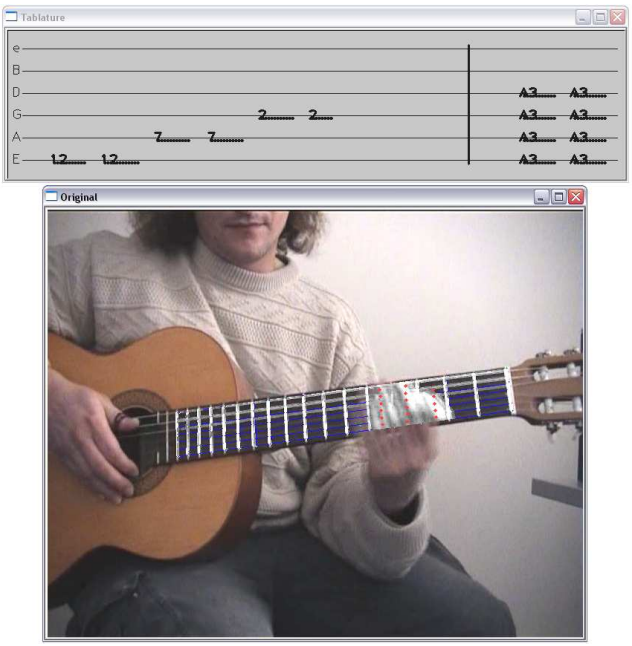


Fig. 2. Interface of the Automatic Transcription System

2. GUITAR TRANSCRIPTION

The typical scenario involved in the discussion of this paper involves one guitarist playing a guitar in front of a web-cam (XviD 640x480 pixels at 25 fps). In the work presented here the entire fretboard of the guitar needs to be completely visible on the video.

2.1. Automatic Fretboard Detection

The first frame of the video is analyzed to detect the guitar and its position. The current version of our system presents few constraints: the guitarist is considered to play a right handed guitar (i.e. the guitar face on the right side) and to trace an angle with the horizontal which does not exceed 90°. The background is assumed to be less textured than the guitar. As a final result, this module returns the coordinates of the corner points defining the position of the guitar fretboard on the video (two outermost points for each detected fret). Guitar frets have some interesting characteristics: they are straight and usually have a different brightness compared to the wood.

The process for obtaining the position of the frets is the following. The Hough transform is employed to find the orientation of the fret board, while the edges are obtained thanks to the Canny algorithm on the original image. The image is then rotated according to the dominant edge orientation in order to align the fret board with the horizontal axis. Wavelet analysis upon the rotated image is performed for enhancing the frets. Then, horizontal projection is performed in order to crop the image to the fretboard only. At this point we have a good estimation of the frets' position but due to some perspective effect the frets may not be straight.

Skewing is applied to the image until the vertical projections are maximized. Candidates (peaks) are chosen on the projection and identified on the original image (by couple). Invalid candidate frets are further filtered out by searching for the maximum energy path between top/bottom and bottom/top extremities. Paths cannot be greater than the distance between the two extremities. Additionally, if the two paths are different then the candidate fret is discarded. At this stage, only valid frets should remain.

2.2. Fretboard Tracking

We have described how the fretboard position is detected on the first frame of the video. We make use of the Tomasi Lukas Kanade algorithm to follow the points along the video.

The coordinates of the end points of each fret are influenced by the movement of the hand. Therefore, some template matching techniques are applied to enforce points to stick to the fretboard. Two constraints were chosen to be invariant to scale, translation or 3D rotations of the guitar: 1) all the points defining the upper (as well as lower) bound of the fretboard must be aligned; 2) the lengths of the frets must comply to the rule $L_i = L_{(i-1)} * 2^{-1/12}$ where L_i represent the length of the i^{th} fret.

To enforce the first constraint a first line is computed that matches the highest possible number of points. The points apart from the line are filtered out and a linear regression (least squares) is computed. All points apart from this sec-

ond line are filtered out and recomputed.

The second constraint is applied by comparing the positions of the points with a template representing the distances of all the frets from the nut (i.e. the fret at the head of the guitar). The best match is found for having the lowest possible number of errors. Points outside the template are removed and their positions are recomputed.

Every twenty seconds the tracking is re initialized to solve any kind of issues which may arise from a wrongful adrifts of the Lukas Kanade point tracking (see section 3). Furthermore, sometimes it may happen that no match can be found because too many points are lost at the same time or because the guitar is not facing the camera. In this cases a new match is searched in the following frames trough the algorithms described in section 2.1.

2.3. Hand Detection

In section 2.2 the methodology employed to follow the position of the frets along the video has been described. Thanks to these coordinates it is possible to separate the region belonging to the fretboard into $n\_strings \times n\_frets$ cells corresponding to each string/fret intersection.

Filtering is done on the frame to detect the skin color and the number of "hand" pixels is counted. A threshold can be applied to detect the presence of the hand (see figure 3.a).

2.4. Audio Visual Information Fusion

Thanks to the audio analysis and standard audio processing techniques [6] we can extrapolate the pitch of the performed notes. This information is converted to a midi file with the information of the note played and the information of the attack time and duration of the note.

For each frame the information about the position of the hand is used to discriminate the correct fret-string couple producing a certain pitch.

Figure 2 shows an example of the developed interface. We can see the interface incorporate two windows. The windows named "Tablature" shows the resulting tablature. The x axis represents the time and the six horizontal lines represent the six strings of the guitar. The vertical line at around 3/4 of the interface represent $t = 0$: at its right the information only comes from the audio analysis; at its left the information is fused together with the video information.

At the right of the line $t = 0$, the same note is represented at the same time on several strings to represent the incertitude that audio brings about when dealing with instruments such as the guitar. Indeed that particular pitch can be played though all the tagged strings.

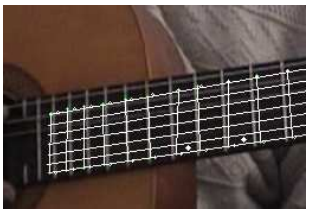
At time 0 (the time represented in the windows named "Original") video is analyzed, the hand is detected at a certain fret and ambiguity is solved. At the left of the line $t = 0$ only one note at time is therefore represented. One may notice that all positions represented in the tablature at the left of the line $t = 0$ generate the same pitch (E3 = $164.81Hz$).

3. PROTOTYPE

We have tested the proposed algorithms on several short videos (around 30 seconds per video). In these videos the guitarist performs different pattern designed to test the algorithms on four different guitars (two classical, one Spanish, and one acoustic). Videos were taken in our laboratories with a DV camera placed on a tripod at less than 2 meters from the guitarist and converted to XviD 640x480 pixels, 25 frames per second at around 250 Kbps. Audio was taken with the integrated camera microphone as well as with a gun zoom microphone to reduce ambient noise.

The guitar tracking algorithm worked correctly all along all the videos. Nevertheless, issues may arise when dealing with fast hand movement which may significantly reduce the number of trackable points and/or slide a consistent number of tracked points in a specific direction. In these cases the two constraints that were described in section 2.2 may not be sufficient to perform a good tracking.

(a) Correct Tracking



(b) Vertical adrift

(c) Horizontal adrift

Fig. 3. Example of video errors.

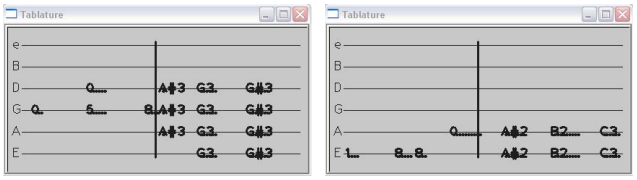
1) *alignment constraint.* If a significative number of points slide up or down the best fitting line may not be exactly parallel to the strings (see figure 3 a).

2) *linear template constraint.* When a significative number of points slide horizontally or it is lost it can happen that the template matching matches better the wrong points than the correct ones. This may result in vertical lines which does not anymore match to the frets borders (see figure 3 b).

With time both these phenomena may be amplified until the tracking is completely lost. We have empirically estimated both these phenomena to be sensible only after 30 to 40 seconds of videos and proceeded to re initialize the tracking

algorithm every twenty seconds using the algorithm described in section 2.

The hand detection was set to detect hand when at least 60% of the cell (i.e. the rectangle defining a fret and a string) contained the hand. This was found to be the minimum percentage allowing to have 0% false positives (which are due to the luminance of frets borders and strings). Setting the threshold at 60% was enough to solve 89% of the note ambiguities (see figure 2).



(a) Ambiguity (b) Fragmentation

Fig. 4. Transcription Errors

In 11% of the cases a note which was played was assigned to two different possible positions. This corresponds to cases in which the played pitch matches with the fundamental pitch of a string (i.e. the pitch the string chime when played without pushing any fret; E2, A2, D3, G3, B3, E4). In this cases both possibilities are actually possible and our system did not disambiguate the note (see figure 4 a).

In around the 3% of the cases one single long note was transcribed as two or more separate notes. This phenomenon was due to the artistic intention of the guitarist who slightly "bended" the string bringing both the hand and the string outside the cell. This will be addressed in future versions of the algorithm (see figure 4 b).

4. FUTURE WORK

A prototype has been described in the former section which demonstrates how the adoption of simple video analysis can help the process of generation of a tablature for guitar music. The example pieces involved in this first prototype only contained a small subset of the possible techniques involved in guitar music. In this section we list some of the possible improvements upon our system.

In the former section we have seen that our system may lose a note when the guitarist "bends" the string. Future work will solve this issue by applying a probabilistic model for the position of the hand. For each cell on the fretboard a $P(h)$ will be computed representing the probability that the hand is both present on the cell and used to play (for example, a part from the case of "barre", only finger tips are used).

Audio analysis will be extended to the polyphonic case allowing for chords and more complex pattern. To help the audio analysis dealing with polyphonic audio we will apply some machine learning techniques to learn prototypical hand positions and shapes (minor chords, major chords and principal variations).

Another system will explicitly perform right hand detection and following to estimate the string attack point to help both the audio and video processing units. Other system may be developed to detect guitar effects such as bending, tapping, slides, hammering on and pulling off, and others.

5. CONCLUSIONS

In this paper we have overviewed a complete, quasi unconstrained, guitar tablature transcription system which uses low cost video cameras to solve string ambiguities in guitar pieces. A prototype was developed as a proof of concept demonstrating the feasibility of the system with today technologies. Results of our studies are positive and encourage further studies on many aspects of guitar playing.

Applications of this research include computer aided pedagogical system which may significantly help guitar students, automatic indexing of song videos through tablature indexing, computer software which may help guitarist create and share music, and many others.

The authors would like to acknowledge the support of the European Commission under contract FP6-027026 K-Space. Some of the ideas reported in this work have been investigated by Valeria Rongione and Lucia Molino under the author's supervision.

6. REFERENCES

[1] A. Burns and M. M. Wanderley, "Visual methods for the retrieval of guitarist fingering," in *NIME '06: Proceedings of the 2006 conference on New interfaces for musical expression*, Paris, France, 2006, pp. 196–199.

[2] J. A. Verner, "Midi Guitar Synthesis: Yesterday, Today and Tomorrow," *Recording Magazine*, vol. 8 (9), pp. 52–57, 1995.

[3] C. Traube, *A Interdisciplinary Study of the Timbre of th Classical Guitar*, Ph.D. thesis, McGill University, 2004.

[4] D. P. Radicioni, L. Anselma, and V. Lombardo, "A Segmentation-Based Prototype to Compute String Instruments Fingering," in *CIM04: Proceedings of the 1st Conference on Interdisciplinary Musicology*, 2004.

[5] B. Zhang, J. Zhu, Y. Wang, and W. K. Leow, "Visual Analysis of Fingering for Pedagogical Violing Transcription," in *MM '07: Proceedings of the 15th international conference on Multimedia*, Augsburg, Germany, 2007, pp. 521 – 524.

[6] Xavier Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, Lisse, the Netherlands, 1997, pp. 91–122.